Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Crowdsourcing Quality Control for Item Ordering Tasks

Toshiko Matsui[1], **Yukino Baba**[1], **Toshihiro Kamishima**[2], **Hisashi Kashima**[1,3]
Univ. Tokyo[1], AIST[2], JST PRESTO[3]

Abstract

One of the biggest challenges in crowdsourcing is quality control which is to expect high quality results from crowd workers who are not necessarily very capable nor motivated. In this paper, we consider item ordering questions, where workers are asked to arrange multiple items in the correct order. We propose a probabilistic generative model of crowd answers by extending a distance-based order model to incorporate worker ability, and give an efficient estimation algorithm.

1 Introduction

One of the most challenging problems in crowdsourcing research is *quality control* to ensure the quality of crowdsourcing results, because there is no guarantee that all workers have sufficient abilities or motivations to complete the offered tasks at a satisfactory level of quality. One promising approaches to this problem is to introduce *redundancy*. We assign a single task to multiple workers, and aggregate their responses by majority voting (Sheng, Provost, and Ipeirotis 2008) or more sophisticated statistical aggregation techniques that consider the characteristics of each worker or task, such as the ability of each worker and the difficulty of each task (Whitehill et al. 2009).

Most of the existing statistical quality control approaches assume that the tasks are binary questions which expect binary answers (e.g., "yes" or "no") or multiple-choice questions. In this paper, we consider *item ordering tasks*, where workers are asked to arrange multiple items in the correct order. We propose a statistical quality control method for item ordering tasks. We model the generative process of a worker response (i.e., an ordering of items) by using a distance-based probabilistic ordering model (Marden 1995). The ability of each worker is naturally incorporated into the concentration parameter of the distance-based model. We also give an efficient algorithm for estimating the true ordering especially when we employ the Spearman distance (Mallows 1957) as the distance measure between two different orderings of items.

We conduct experiments using word ordering tasks and sentence ordering tasks by using a commercial crowdsourcing marketplace. We compare our quality control method to an aggregation method which does not consider the abilities of workers. The experimental results show that our method achieves more accurate answers than the baseline method.

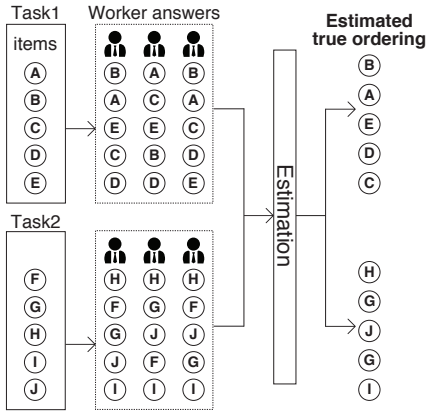


Figure 1: Overview of quality control problem for item ordering tasks in crowdsourcing. Our goal is to estimate true ordering of items from crowd-generated answers.

2 Crowdsourcing Quality Control Problem for Item Ordering Tasks

Let us assume that we have I ordering tasks, whose i-th task has M_i items to be ordered. We represent the true order as a *rank vector* $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,M_i})$, where $\pi_{i,j}$ indicates the position of item j of task i in the true order of the items of M_i (Marden 1995). For example, if we have a task with five items indexed as $1, 2, 3, 4,$ and 5, and the true order of them is given as $(2, 4, 1, 3, 5)$, then the true rank vector is $(3, 1, 4, 2, 5)$. Note that $\boldsymbol{\pi}_i$ is a permutation of $(1, 2, \ldots, M_i)$. We resort to crowdsourcing to obtain estimates for the true rank vectors. We assume that we employ K crowdworkers in total. We denote by $\mathcal{I}^{(k)}$ the indices of tasks that the k-th worker work on, and by $\mathcal{K}_i$ the indices of workers who work on the i-th task. We also denote by $\boldsymbol{\pi}_i^{(k)} = (\pi_{i,1}^{(k)}, \pi_{i,2}^{(k)}, \ldots, \pi_{i,M_i}^{(k)})$ the rank vector that the k-th worker gives to the i-th item ordering task. Our goal is to estimate the true rank vectors $\{\boldsymbol{\pi}_i\}_{i\in\{1,2,\ldots,I\}}$ given the (unreliable) rank vectors $\{\boldsymbol{\pi}_i^{(k)}\}_{k\in\{1,2,\ldots,K\},i\in\mathcal{I}^{(k)}}$ collected by using crowdsourcing.

3 Proposed Method

To solve the problem of aggregating the crowd-generated answers for item ordering tasks, we first give a statistical model of the generative process of worker responses. Our model is based on a distance-based ordering model (Marden 1995). The probability of a rank vector $\tilde{\boldsymbol{\pi}}$ given a true order $\boldsymbol{\pi}$ and the k-th worker's ability parameter $\lambda^{(k)} > 0$ is defined as

$$\Pr[\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}, \lambda^{(k)}] = \frac{1}{Z(\lambda^{(k)})} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right),$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance (also known as the *Spearman distance* in the ranking model literature), and $Z(\lambda^{(k)})$ is a normalizing constant.

We apply maximum likelihood estimation to obtain estimates for the true rank vectors as well as the worker ability parameters. The log-likelihood function L is given as

$$L(\{\lambda^{(k)}\}_k, \{\boldsymbol{\pi}_i\}_i) = -\sum_k \sum_{i \in \mathcal{I}^{(k)}} \Bigg\{ \lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i) + \log \sum_{\tilde{\boldsymbol{\pi}}} \exp\Big(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_i)\Big) \Bigg\}. \quad (1)$$

Our strategy to optimize the objective function (1) w.r.t. $\{\lambda^{(k)}\}_k$ and $\{\boldsymbol{\pi}_i\}_i$ is to repeat the two optimization steps: the one w.r.t. $\{\lambda^{(k)}\}_k$ and the one w.r.t. $\{\boldsymbol{\pi}_i\}_i$. Given all the worker ability parameters $\{\lambda^{(k)}\}_k$ fixed, the optimal true rank vector $\boldsymbol{\pi}_i$ for task i is given as follows. First, for each item $m (= 1, \ldots, M_i)$, we calculate a *weighted rank* $w_{i,m}$ which is a weighted mean of the ranks given by workers weighted by the worker abilities, that is,

$$w_{i,m} = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \lambda^{(k)} \pi_{i,m}^{(k)}.$$

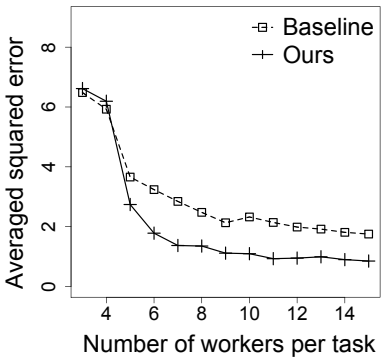
The true item ordering is given by sorting the items by $w_{i,1}, w_{i,2}, \ldots, w_{i,M_i}$ in ascending order.

Optimization with respect to the worker ability parameters $\lambda^{(k)}$ with true rank vectors $\{\boldsymbol{\pi}_i\}_i$ fixed is easily performed by numerical optimization, since the objective function (1) is represented as the sum of K independent optimization problem with only one variable.

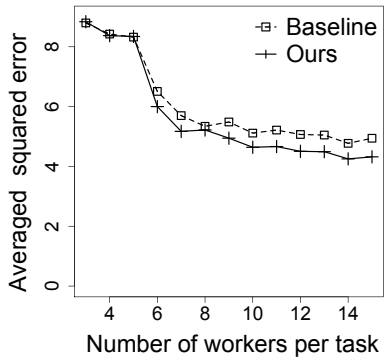
4 Experiments

We used a crowd sourcing marketplace Lancers (http://lancers.jp) to collect two crowdsourced datasets; one is for word ordering tasks, and the other is for sentence ordering tasks. Word ordering is a task aiming to order given English words into a grammatically correct sentence. Workers are shown an English sentence with five or six randomly shuffled words, and they are asked to correct the order of the words. Sentence ordering is a task to order given sentences so that the aligned texts logically make sense. Workers are presented a paragraph consisting of five or six sentences whose order is permuted, and they are requested to rearrange them correctly.

We applied our method to the two crowd-generated datasets, and calculated the Spearman distance (i.e., the squared error) between each estimated rank vector and the ground truth rank vector. We also tested a baseline method which does not consider worker ability. In order to investigate the impact on the estimation accuracy by the number of workers assigned to each task, we randomly selected n (ranging from 3 to 15) workers from the all workers for each task, and only used the selected responses for estimation. We examined the averaged estimation errors of 50 trials. Figure 2 shows that the proposed method outperformed the baseline method for more than five or six workers.



(a) Word ordering



(b) Sentence ordering

Figure 2: Experimental results.

References

Mallows, C. L. 1957. Non-null ranking models. I. *Biometrika* 44:114–130.

Marden, J. I. 1995. *Analyzing and Modeling Rank Data*, volume 64. CRC Press.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *ACM SIGKDD*.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*.